Reasoning by equivalence:The potential contribution of an automatic proof checker

Reasoning by equivalence as elementary formal proof: the potential contribution of an automatic proof checker

Christopher Sangwin

June 28, 2018

Abstract

Reasoning by equivalence, a form of line-by-line algebraic reasoning, is the most important single form of reasoning in school mathematics. In this chapter I define reasoning by equivalence and examine the role of reasoning by equivalence in mathematical proof. I base the discussion on an examination of the extent to which students are currently asked to "prove", "show" or "justify" in high-stakes national examinations. I then report research into how students go about solving such problems on paper. These results inform the design of an automatic proof checker within the STACK software which assesses students' responses. I report on the use of this software with students. Finally I discuss the implications of this work for what constitutes mathematical "proof" at school level, and how this might be taught and learned online.

keywords: automatic assessment; proof checker; algebra; mathematics education.

1 Introduction

The heart of mathematics is setting up abstract problems and solving them, and the outcome of this process is a "proof". Polya (1962) identified four patterns of thought to help structure thinking about solving mathematical problems. His *"pattern of two loci"* is highly geometric. To use the pattern of two loci keep only one of the conditions needed and solve this simpler problem, expecting to get a set (locus) of solutions in this less constrained situation. Keep each condition in turn, and the overall solution is where the loci intersect. For example, when finding the tangents to a circle centred at O through an external point A, the tangent line must be at right angles to the radius. The set of points P for which OP is perpendicular to AP is the circle centered at the midpoint of OA through O. This is one locus, and the circle itself is the second. The *"superposition"* pattern also breaks constraints into separate parts in a more algebraic way. Examples include Lagrange interpolation, and solving linear differential equations. The *"recursion"* pattern solves a problem by using a smaller (or simpler) case, for example finding the binomial coefficients using Pascal's triangle.

Legitimate patterns of thought directly translate into what is considered to be an acceptable proof. For example, recursion solves the problem and proof by induction is the resulting formal justification.

The *"Cartesian"* pattern is where a problem is turned into a system of equations, before the equations are solved using algebra. Note that the algebraic manipulation is the technical middle step in the process: setting up the equations and interpreting the solutions are essential parts to complete this pattern. My previous work (Sangwin & Köcher, 2016) examined questions set in school-level examination papers and found that line-by-line algebraic reasoning, termed by Nicaud, Bouhineau, and Chaachoua (2004) as *reasoning by equivalence*, is the most important

single form of reasoning in school mathematics. This is closest to Polya's *"Cartesian pattern"* and this predominates normative answers to school examination questions.

Reasoning by equivalence is essentially a symbol-pushing technique, which reduces algebraic reasoning to a mechanical calculation. There is nothing pejorative in describing this activity as symbol-pushing or as mechanical: in some senses this automation is liberating. Indeed Leibniz (1966) sought a "universal calculus" and Boole (1847) explicitly sought to replace syllogisms in language with symbolic calculus. This work continues with attempts to automate proof more generally, see Beeson (2004) for a survey.

This research is based on the following epistemological position: to successfully automate a process it is necessary to understand it profoundly. It follows that automation of a process necessitates the development of a certain kind of understanding. I am trying to automate the assessment of students' line by line algebraic reasoning, including provision of effective feedback on their progress. As I hope to show, the attempt to implement a line by line assessment system reveals ambiguities, inconsistencies and outright errors in elementary algebra as currently taught and learned. For example, using algebraic rules uncritically outside the domains of definition such as $\sqrt{ab} = \sqrt{a}\sqrt{b}$ leads to contradictions such as $-1 = 1$, which have been examined elsewhere, e.g. (Bernardo & Carmen, 2009), (Tirosh & Evan, 1997) and (Levenson, 2012). While professionals might dismiss such errors as trivial they have been made by the very best mathematicians in the past (e.g. see (Euler, 1822, p. 43, §148)) and students regularly continue to do so, (Kirshner & Awtry, 2004).

In this chapter I examine the role of reasoning by equivalence in mathematical proof. This chapter contains a number of sections. First I define the mathematics of reasoning by equivalence. By looking at what students are asked to do in high-stakes national examinations I examine the extent to which students are currently asked to "prove", "show" or "justify" and the mathematics this involves. I then report research into how students go about solving such problems on paper. This evidence has been used to inform the design of software which assesses students' responses, and I report on the use of such software with students. I then discuss the implications for what constitutes mathematical "proof" at school level, and how this might be taught and learned online.

2 Reasoning by equivalence

In educational terms reasoning by equivalence is a loose collection of rules, such as "doing the same thing" to both sides of an equation. In Sangwin and Köcher (2016) we did not define the process more formally, indeed the rules of elementary algebra are not normally articulated carefully. When seeking to develop technology, a loose collection of rules is simply not sufficient or satisfactory and so in this section I define reasoning by equivalence.

The phrase "doing the same thing" to both sides of an equation has a focus on the individual steps, and the legitimacy of working in steps. Reasoning by equivalence does not follow a prescribed routine, with fixed size steps in working. Instead, the move between adjacent lines from one expression to the next is legitimate if and only if adjacent expressions are equivalent. In most situations many algebraic "steps" are combined, including associativity, commutativity and basic integer arithmetic. Hence, defining what is and is not a legitimate step turns out to be problematic in an educational context.

Let $X \subset \mathbb{K}$ where $\mathbb{K}$ is a set of numbers given by the context, such as $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$. For the purposes of this chapter, I define "a correct argument" as an ordered list of mathematical expressions $E_1, \cdots, E_n$ so that all the E_j are equivalent. For example, in the case of equations in a single variable x this means the solution set of E_j is precisely the same for all $j = 1, \cdots, n$. The role of $\mathbb{K}$ is in deciding what is and is not a solution, e.g. $x^4 - 9 \equiv x^2 - 3$ in $\mathbb{R}$, but not in $\mathbb{Q}$ or $\mathbb{C}$. Reasoning by equivalence is the process by which an individual takes successive representatives from a particular equivalence class.

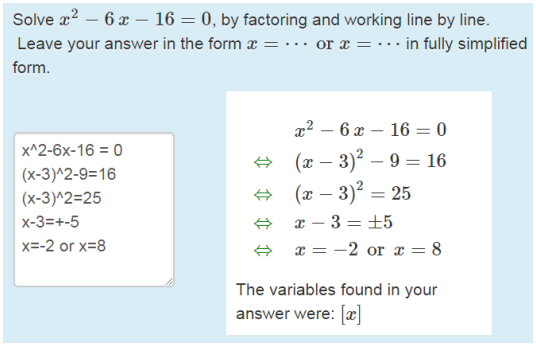

Figure 1: Reasoning by equivalence in the STACK system

The advantage of defining the correctness of an argument in terms of equivalence is that the student is free to take any route they choose. In an educational context, overall correctness will additionally require that E_1 is some specific expression (i.e. the problem to be solved) and that E_n is given as a specific form, such as $x =?$ in the case of a linear equation. A teacher might also require some other properties from a student for an answer to be fully correct. An example of a student solving a quadratic equation using reasoning by equivalence, as implemented in the STACK system (see Sangwin, 2013), is shown in Figure 1. In this case, the student has correctly solved the quadratic specified and the (automatically generated) $\Leftrightarrow$ symbols[1] indicate the equivalence of adjacent lines. However, the final system-generated feedback (omitted in the figure) actually says *"The question asked you to solve the equation by factoring. The factored form should appear as one line in your working."* and no numerical marks are awarded: the student has not followed the instructions in the question. This example is typical of many educational situations. "Correctness" is a combination of properties, and correct line by line reasoning is only one of them.

Two expressions are equivalent if they take the same value when evaluated at each of the points in the domain of the variables. For example, $x^2 - 1$ is equivalent to $(x-1)(x+1)$ on the domain of real numbers. The difference between equations and expressions is illustrated by the following example.

$$|x - 1/2| + |x + 1/2| - 2 = 0 \Leftrightarrow |x| - 1 = 0$$

but

$$|x| - 1 \neq |x - 1/2| + |x + 1/2| - 2 \text{ for all } x \in \mathbb{R}.$$

I shall write $\equiv$ to indicate equality of expressions on their domain of definition, the third bar indicating the stronger condition. That is

$$p \equiv q \Leftrightarrow p(x) = q(x), \quad \forall x \in X.$$

Equivalence of equations can be defined in two useful ways, and for educational purposes both are needed. $V(p) = \{x \in X | p(x) = 0\}$ is called the *variety* defined by p. This is a geometric notion of the set of solutions, but the variety lacks any way to distinguish between repeated roots. Formally, and for completeness, for systems of polynomial equations the multiplicity of the roots can be distinguished by calculating the reduced Gröbner basis, with respect to a specified order of the variables which gives a canonical representation of the equivalence class, see Adams and Loustaunau (1994). For a set of polynomials in a single variable the Gröbner basis is given by the highest common factor, and hence provides the mutual solutions including multiplicity. For a set

[1]The symbol $\Leftrightarrow$ should be read "if and only if" just as $=$ is read "is equals to". The symbol $:=$ is read as "is defined to be".

of linear equations in many variables the Gröbner basis is equivalent to the reduced echelon form, with respect to an ordering of the variables.

These abstract definitions do not help someone learn *how to solve* algebraic problems. The issue then is the algebraic operations applied to the expressions on both sides of an equation that provide an equivalent equation. For the purposes of this discussion, I restrict to single variable expressions over the set $X \subset \mathbb{K}$, the natural domain of the expressions. Solving the equation $p = q$ necessitates finding those values of $x \in X$ so that $p(x) = q(x)$. To legitimately do the same thing to both sides of an equation $p = q$, it is necessary to ensure that

$$p = q \Leftrightarrow f(p) = f(q).$$

Since f is a function

$$p = q \Rightarrow f(p) = f(q).$$

A function $f\colon X \to Y$ is *injective* if

$$\forall p, q \in X, f(p) = f(q) \Rightarrow p = q.$$

Hence "doing the same thing to both sides" is legitimate if and only if f is an injective function.

Note that multiplication by a term a is an injection if and only if $a \neq 0$. Cases when a is an algebraic term which might be zero give rise to various false arguments, see e.g. (Maxwell, 1959) and (Northrop, 1945). This problem can be entirely avoided by *auditing* to track the side condition $a \neq 0$, see Sangwin (2015).

A substitution[2] is a syntactic transformation of a formal expression in which a variable, sub-expression, or term, is consistently replaced by other expressions. For substitution assume $s\colon X \to X$ is a function, then substitution involves replacing a term x by $s(x)$. For this to be legitimate it is necessary to ensure that

$$p(x) = q(x) \Leftrightarrow p(s(x)) = q(s(x)), \quad \forall x \in X.$$

Since s is a function it follows immediately that

$$p(x) = q(x) \Rightarrow p(s(x)) = q(s(x)), \quad \forall\, x \in X.$$

However in general the converse is false. For example, let $X = \mathbb{R}$, $p(x) := x$ and $q(x) := |x|$ and $s(x) := x^2$. Then $x^2 = |x^2|$ but $x \neq |x|$ for all $x \in \mathbb{R}$. If $s\colon X \to X$ is surjective then

$$\forall x' \in X, \exists x \in X \text{ such that } x' = s(x).$$

The hypothesis $p \circ s \equiv q \circ s$ and the assumption of surjectivity of s prevents a counter example of the form $\exists x' \in X\, p(x') \neq q(x')$.

In summary, "doing the same thing to both sides" retains equivalence if and only if f is injective, and substitution retains equivalence if and only if s is surjective. In both these definitions the domain X is crucial.

I should acknowledge that since there is no conscious "direction of travel" or sense of "progress", in educational terms, reasoning by equivalence looks very odd. It is entirely possible for correct reasoning to have repetitive steps, unnecessary loops or digressions. Aesthetic judgements are a separate matter from correctness, as is what might be considered an appropriate "jump" or level of detail. Only further experience and development will allow us to decide if additional aesthetic

[2]Note, by "substitution" I mean capture-avoiding substitution. A substitution is said to be a *capture-avoiding substitution* when the process avoids accidentally allowing free variables in the substitution to be captured inside the original expression. For example, in the function $x \mapsto x\,y$ if I replace y with x, the function would become $x \mapsto x \times x$ which is different. Both xs now refer to the argument of the function. The second x in $x \mapsto x \times x$ which was originally "free" has been "captured".

	M	A	R	N
# of marks	65	50	3	12
%	50	38	2	9
IB %	31	50	4	15

[M] Marks awarded for attempting to use a correct Method; working must be seen.
[A] Marks awarded for an Answer or for Accuracy: often dependent on preceding M marks.
[R] Marks awarded for clear Reasoning.
[N] Marks awarded for correct answers if no working shown.

Table 1: Descriptive statistics of marks available on Higher Mathematics questions from 2015.

measures can be automated. For example, an aesthetic measures might include "distance from a model answer".

A more serious, and unfortunate, consequence of this definition is that some correct equivalence reasoning arguments do not correspond to correct mathematical steps.

$$\begin{aligned} & x^2 - 6x + 9 = 0 \\ \Leftrightarrow \quad & (x-5)(x-1) = -2 \times 2 \\ \Leftrightarrow \quad & x - 5 = -2 \,\text{or}\, x - 1 = 2 \\ \Leftrightarrow \quad & x = 3 \end{aligned}$$

Fortunately good nonsense of this kind is surprisingly hard to find.

3 Reasoning and school examinations

My previous joint work, (Sangwin & Köcher, 2016), examined questions set in school-level examination papers and found that a third of the marks for school examinations were awarded for reasoning by equivalence. In Sangwin and Köcher (2016) our methodology was to take a corpus of published examination questions, together with the official mark scheme. We examined the extent to which we could automatically mark answers to these questions using the STACK software in a way which was faithful to the published mark scheme. In that research we selected the specimen questions on paper 1 and paper 2 for International Baccalaureate[3] (IB) Mathematics Higher level, for first examinations in 2008. In this section I repeat the analysis, using different questions and with the benefit of a number of years of software development. The reasoning by equivalence engine now exists as a working prototype, and this research provides an opportunity to test my previous claims regarding the potential automation of assessment of students' equivalence reasoning.

The IB specification has now been superseded and so for this chapter I took the two Mathematics papers from the 2015 Scottish Higher Mathematics examinations. In Scotland school students typically take five Higher subjects aged 16–17, which form the basis of university entrance criteria. Universities outwith Scotland, e.g. in the rest of the UK, may require students to study to Advanced Higher level aged 17–18, so that the Highers do not perfectly align with the IB. Paper 1 is a non-calculator paper with a 70 minute time duration, and with 60 marks available. Paper 2 permitted the use of a calculator, with a 90 minute time duration, and with 70 marks available. Students sat both papers on the morning of Wed 20 May 2015. The examination papers, together with the official mark scheme, are publicly available from the Scottish Qualifications Authority website.

For the purposes of this chapter I repeated the following methodology from Sangwin and Köcher (2016). The IH and Highers examination systems are similar, but not identical. However the similarity and the specificity of the supplied mark scheme made it is straightforward to map the Higher marks onto the IB classifications, providing comparability with the previous work.

[3]International Baccalaureate is a registered trademark of the International Baccalaureate Organization.

	# marks	
(i) Awarded by STACK *exactly*	47	36%
(ii) Of which reasoning by equivalence	35	27%

Table 2: The extent to which Higher Mathematics questions can be automatically assessed

My analysis of these questions sought to determine the extent to which answers could be automatically assessed by STACK exactly as in the mark scheme. I noted the extent to which this assessment included reasoning by equivalence as the method marks. The results are shown in Table 2. For the Highers, I implemented 27% of the marks as reasoning by equivalence. I previously suggested that for the IB papers 36% of the marks could be awarded. This discrepancy can be accounted for, since there is still reasoning by equivalence which is not yet accessible because calculus operations are not currently included. Typically, students either algebraically rearrange an expression before integrating/differentiating (e.g. see Highers paper 1 Q7, and paper 2, Q8b) or they perform calculus and rearrange the result, or form an equation. Inclusion of calculus operations in STACK would substantially increase the range of answers which could be fully assessed. These results, therefore, represent a lower bound on the extent to which school examinations could ultimately be automatically assessed.

As with the IB, the "reasoning" marks are an almost insignificant component of the examination marks. In the case of the Highers questions I decided that only 3 marks (from 130) were available for reasoning. Higher paper 1 Q3 asked students to use the remainder theorem to show $(x+3)$ was a factor of $x^3-3x^2-10x+24$; Higher paper 1 Q9 required students to establish if points were colinear; and paper 2 Q3b asked students to reason about the model based on derived inequalities. Most questions asked students to calculate, rather than reason based on calculations.

The results reported in this section support the hypothesis that the most important single form of reasoning in school mathematics is, and remains, algebraic reasoning by equivalence. In many cases students' first, very simple, proofs are entirely algebraic. Even where additional reasoning is used, reasoning by equivalence is a central component of most current school-level proofs. In the next section I describe students' traditional paper and pencil attempts at solving algebraic problems, before reporting on their attempts at reasoning by equivalence with the STACK software.

4 Students' attempts at algebraic reasoning

Following from the research reported in Sangwin and Köcher (2016) and curious to pursue the potential of reasoning by equivalence as a starting point for assessment of complete mathematical arguments, albeit of the most elementary kind of proof, I started to consider how software might automatically assess the correctness of a student's algebraic derivation. To inform this design process I investigated students' attempts at algebraic reasoning. My prior experience as a teacher strongly suggests that students' written algebraic work typically contains no logical connectives or other justification and entirely ignores natural domain conventions. When solving an equation students appear to work line-by-line, but each line is apparently disconnected from the previous lines. There is some literature on this issue, e.g. (Bernardo & Carmen, 2009), (Levenson, 2012), (Tirosh & Evan, 1997), but none of this directly relates to university mathematics students. I therefore set out to investigate the following, working on paper in the traditional way.

1. To what extent do students use (and correctly use) logical connectives between lines of algebraic working? What other justification is evident other than "implied equivalence"?
2. To what extent do students acknowledge the natural domains of definition? For example in the expression $\frac{1}{x-4}$ the value $x=4$ is excluded from the natural domain.

$$\begin{aligned}
\frac{x+5}{x-7} - 5 &= \frac{4x-40}{13-x} \\
\frac{x+5-5(x-7)}{x-7} &= \frac{4x-40}{13-x} \\
\frac{4x-40}{7-x} &= \frac{4x-40}{13-x} \\
7-x &= 13-x \\
7 &= 13.
\end{aligned}$$

Figure 2: An erroneous solution to question (1)

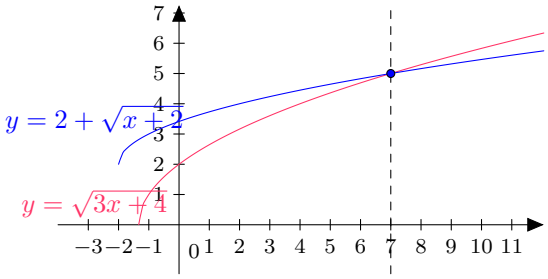

Figure 3: A graphical solution to question (2)

3. To what extent do students show evidence they have checked a particular answer is actually correct?

In Sangwin (2016) my experimental instrument to address the above research questions consisted of the following two algebraic problems.

$$\text{Q1. Please solve } \frac{x+5}{x-7} - 5 = \frac{4x-40}{13-x}. \tag{1}$$

$$\text{Q2. Please solve } \sqrt{3x+4} = 2 + \sqrt{x+2}. \tag{2}$$

These two question were chosen because they require line by line reasoning but there are opportunities for applying rules outside the domain of applicability. For example, an erroneous solution to the first question, taken from Northrop (1945, p. 81), is shown in Figure 2. The "rule" $\frac{a}{b} = \frac{a}{c} \Leftrightarrow b = c$ obscures the requirement that $a \neq 0$. The purpose of including this question was to see how many students cancel a term such as $4x - 40$, and evidence for how students dealt with the subsequent contradiction.

The typical algebraic solution to the second question, taken from Newman and et.al. (1957, p. 8), involves squaring both sides and collecting terms to get $x - 1 = 2\sqrt{x+2}$ and squaring again to obtain a quadratic with real roots $x = 7$ or $x = -1$. However, there is only one real solution as shown in the graphical solution of Figure 3. It is relatively straightforward to construct similar examples with no real solutions (e.g. $\sqrt{x+2} = 2 + \sqrt{3x+4}$), see (Bonnycastle, 1836, p. 88), (Durell, 1930, p. 46) or (Lund, 1852, p. 130). In some problems of this kind the natural domain can be used to eliminate one of the solutions, but $x = -1$ does not violate the domain constraints of $\sqrt{3x+4}$ or $\sqrt{x+2}$, i.e. $x \geq -4/3$. In this case it the reversibility of squaring both sides of an equation which introduces a spurious solution, rather than domain constraints.

As reported in Sangwin (2016), students solved these two equations writing answers on paper. The cohort were a group of 175 students taking an engineering programme at a good United Kingdom

university. The detailed methodology and results are reported in Sangwin (2016). For question 1, of the 113 correct responses, 53 (46.9%) cross multiplied, expanded out all brackets and solved the resulting equation correctly to get the unique answer $x = 10$. Therefore, these students missed the opportunity to cancel the term $4x - 40$ on both sides. In this question 25 (22%) students started by writing the left hand side as a rational expression. While 22 of these students had the opportunity to cancel a factor none of them did so. For question 1, only 14 (9.5%) of students showed any evidence of logical connectives between algebraic statements. Only 2 students wrote any evidence of having performed a check that their answer satisfied the original equation, and only 1 student explicitly considered domains of definition of the rational expression by excluding $x = 7$ and $x = 13$ from the domain of definition for the original equation. Overall, only 17 (11.6%) of students wrote any evidence of more than algebraic symbolic manipulation.

For question 2, the most popular answer consisted of squaring both sides, rearranging and squaring again before solving the resulting quadratic to get roots $x = 7$, $x = -1$. 85 students took this approach, of which 24 students also checked that their answers satisfied the original equation, giving complete and correct solutions by only 16% of the cohort. For question 2, only 4 students showed any evidence of checking domains of definition and only a further 3 students used any logical connectives. The most common mistake was squaring a binomial incorrectly, e.g. $(\sqrt{a} + \sqrt{b})^2 = a + b$. In particular, 18 students wrote

$$\sqrt{3x+4} = 2 + \sqrt{x+2}, \quad 3x + 4 = 4 + x + 2, \quad x = 1. \tag{3}$$

What was perhaps surprising was that even for comparatively elementary problems, students took on average 10 or 14 lines to achieve a correct solution to questions (1) & (2) respectively. This strongly suggests that online assessment systems, such as STACK, need to assess more than the final answer, particularly in a formative setting.

5 Developing algebraic reasoning by equivalence in STACK

Based on the analysis of examination questions, and research associated with students' algebraic reasoning, I started to develop an extension to the STACK online assessment system to assess students' reasoning by equivalence. Initially I decided that the minimum worthwhile functionality should include (i) rearranging a simple equation to make a particular variable the subject, and (ii) solving linear and quadratic equations in a single real variable, over the real numbers. However, ultimately the first version included sufficient additional functionality to allow automatic assessment of the problems in questions (1) and (2).

In developing a reasoning by equivalence system I adopted the following design principles.

- P1 The system should be mathematically and logically correct.
- P2 Students should provide a complete line by line solution. A student's answer constitutes a single mathematical object, which is to be subject to verification.
- P3 The system should mirror current practice as closely as possible, within the constraints of a liberal typed linear syntax, see Sangwin and Ramsden (2007).

A distinctive aspect of this design is the focus on the whole argument as a single object which can be subject to formal verification. That is to say, the student's lines of working are assumed to be connected by equivalence symbols and this becomes the single mathematical object. Similarly, I decided that students should not need to explicitly add in the natural domain of expressions. However, in line with (P1), the system would indicate natural domain constraints automatically.

There is an important distinction between reasoning and argumentation.

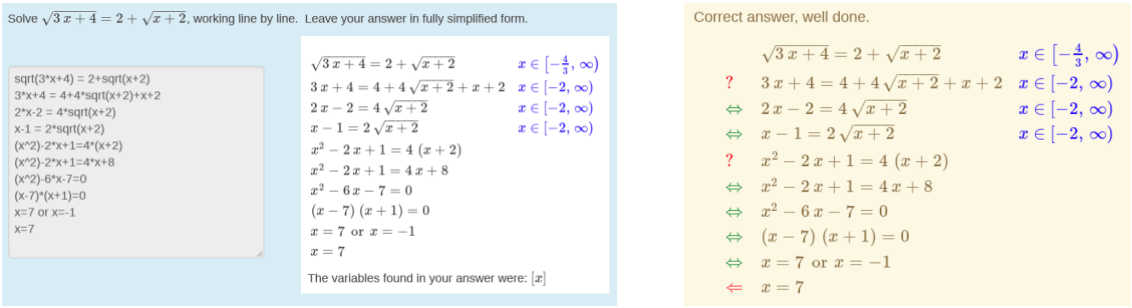

Figure 4: A student's attempt at reasoning by equivalence in the STACK system

> *Reasoning* is [...] the line of thought adopted to produce assertions and reach conclusions. *Argumentation* is the substantiation, the part of the reasoning that aims at convincing oneself or someone else that the reasoning is appropriate. (Boesen, Lithner, & Palm, 2010, p. 92)

Therefore reasoning can be valid or invalid. Similarly, argumentation may or may not correctly justify a particular step. Teachers often use a phrase such as "right method, wrong reason" to describe various types of mistake. In line with (P3) I decided that, in the first version, students would only be required to provide their reasoning (i.e. the algebraic expressions) but *not* also supply the argumentation to justify what they have done or why.

The decision not to require students to provide details of (i) argumentation and (ii) natural domains significantly simplified the input mechanism. STACK already has a well-developed input mechanism for algebraic expressions, as described in Sangwin and Ramsden (2007) and later in Sangwin (2013). A distinctive feature of this interface is a separation of "validity" of input from "correctness" of an answer. Feedback relating to the validity of input includes syntactic problems such as missing brackets, and this feedback is intended to always be available even during an examination or other high-stakes situation. The provision of validity feedback mediates the difference between a typed expression (e.g. `1/(1+x^2)`) and traditional two dimensional displayed forms (e.g. $\frac{1}{1+x^2}$), helping students to ensure the expression they typed matches what they intended. Rejecting expressions as "invalid" rather than "wrong" immediately helps prevent students from being penalised on a technicality. Separating validity from correctness has been found to be an essential feature of effective online assessment of mathematics.

The interface, implementing question 2, is shown in the left of Figure 4. The student has typed a number of lines of working into the textarea on the left. STACK has automatically and instantaneously displayed their line by line working in the box to the right. In addition to the expressions typed by the student, STACK has inferred and indicated the natural domain as blue text to the right. The right hand side of Figure 4 shows the outcome of the assessment. In this case, the line by line reasoning is not correct in three places. Two involve squaring both sides (and additionally a domain enlargement), whereas the last line omits a root from the previous line without justification. Here the system uses a $\Leftarrow$ symbol to indicate the subset of one variety within another.

Notice that in Figure 1 STACK added in $\Leftrightarrow$ symbols into the validation feedback area to indicate lines are equivalent, whereas these have not been added automatically in the left of Figure 4. In some situations a teacher may want to immediately indicate whether adjacent lines are equivalent. In other situations feedback on whether lines are equivalent constitutes part of what makes up "correctness" of the answer, and so the equivalence symbols are not displayed until the student considers they have a complete and correct solution worthy of assessment. The teacher is faced with a large number of choices about the nature and timing of feedback, and what should be part

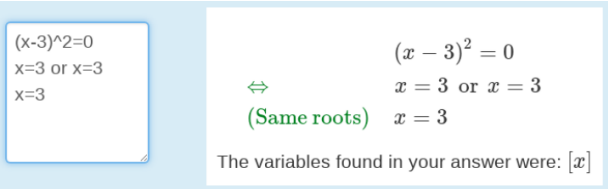

Figure 5: Dealing with repeated roots

of the validity check, and what is core to the correctness of an answer.

The interface was developed by me, and question 2 was trialled on a year 1 calculus course at a good UK university. The course contains 568 students, academically comparable to the group who undertook the original paper and pencil trial (but at different institutions, and some three years later). In this situation students were permitted multiple attempts at the quiz, and Figure 4 shows one student's attempt. Over 477 students provided valid attempts to this question. Of these attempts 33 students only stated the question (valid, but wrong). A further 33 students stated the question and jumped to the final answer $x = 7$ (valid, and correct) and five students stated the question and gave $x = 1$ as the next line (valid, but wrong, and note (3)). If a teacher wants more than just the question and a final answer an additional property, such as the minimum number of lines, would need to be specified. There were 387 different valid responses, with only 15 being given more than once. What is interesting is the number of lines of working used in students' final answers. Of the students who did more than state the question and answer, i.e. who provided three or more lines of working, the mean number of lines used was 8.0 with a standard deviation of 2.4 and only 12 students used more than 12 lines. In the online interface students used fewer lines than for the same problem on paper. The ability to copy, paste and delete input lines within the textarea makes it likely students tidied up their answer before final submission, reducing unnecessary lines.

The original goal was to implement minimum functionality to support (i) rearranging a simple equation and (ii) solving linear and quadratic equations. In order to achieve this functionality the following issues needed to be addressed.

- There is general ambiguity about how to express multiplicity of roots. If $(x-1)^2 = 0$ is not equivalent to $x = 1$ then students need to indicate multiplicity of roots, but I am aware of no consensus on how this should be notated.
- Students need to enter sets of real numbers, sometimes as a set, sometimes using interval notation and at other times via inequalities.

My approach to multiplicity of roots is illustrated in Figure 5. The equation $(x-3)^2 = 0$ and the expression "$x = 3$ or $x = 3$" are considered to be equivalent, because they have the same roots with the same multiplicity. The expressions "$x = 3$ or $x = 3$" and "$x = 3$" have the same variety, but are not identical. This is, of course, slightly awkward since logical "or" is idempotent, and so "$x = 3$ or $x = 3$" and "$x = 3$" would be equivalent at a symbolic level. For this reason, STACK accepts $x = 3$ as equivalent to $(x-3)^2 = 0$, but with an acknowledgement.

This leads us onto the issue of entry of sets of real numbers. If reasoning by equivalence is merely choosing a list of representatives of an equivalence class, then any of these representatives describes the set of real numbers! In an educational context, teachers normally look for more normative ways of writing this set, such as $x = 1$. After much thought, and experimentation, I decided to *require* forms such as "$x = 1$ or $x = 2$" rather than the alternatives "$x = 1, 2$" or "$x = 1$ or 2". The statement "$x = 1$ and $x = 2$" represents no real numbers (i.e. the empty set) and so is normally simply wrong, although the sentence "The roots are 1 and 2." is perfectly correct and

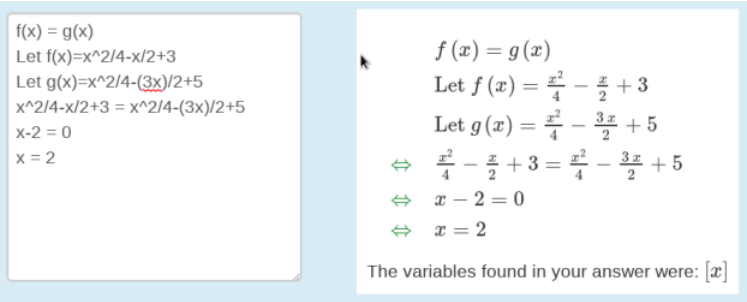

Figure 6: Support for assignment of a value to an expression within an argument

is in no way ambiguous. Currently I do not support entry of expressions like $x \in \{1, 2\}$ although this will be added in a future version, as will be support for interval notation.

What became clear during the automation of questions from the Scottish Highers Examinations, was that the addition of a small number of additional mathematical moves, combined with reasoning by equivalence, would significantly expand the range of mathematical questions which can be completely assessed automatically. In particular (i) a "let" assigning a value to an expression, (ii) support for equating coefficients, (iii) support for basic symbolic calculus operations.

As an example of the "let" statement consider the response shown Figure 6, taken from Scottish Highers (2015), paper 2, Q4a. This does not, strictly speaking, maintain an equivalence but it proves very useful indeed if the system will allow evaluation of an expression, followed by subsequent algebraic manipulation. Beyond the scope of this chapter is discussion of support for basic symbolic calculus operations, which the STACK system already supports extensive functionality to implement. Suffice to say, that differentiating an expression, equating to zero and solving the subsequent equation is a very common task at this level.

Designers of other tutor software have taken very different decisions. For example, Mathpert is a stand-alone desktop system which allows its user to solve mathematical problems by constructing a step-by-step solution.

> *Mathpert* is intended to replace paper-and-pencil homework in algebra, trig, and calculus, retaining compatibility with the existing curriculum while at the same time supporting innovative curriculum changes; to provide easy-to-use computer graphics for classroom demonstration in those subjects, as well as for home study; to replace or supplement chalk-and-blackboard in the classroom for symbolic problems as well as graphs. (Beeson, 1998)

Students either pick a built-in topic or enter the problem they wish to solve. Students then use the *calculation window* to solve the problem in a step-by-step fashion. To do this, users select part (or all) of an expression and the system responds with a menu of operations which can be performed on that selection. The software performs the selected operation automatically. Having written my own software, I full endorse its fundamental approach to the mathematical underpinning.

> ... if we start with an *educational* purpose, [...] it is impossible to achieve ideal results by tacking on some additional "interface" features to a previously existing computational system. To put it another way: it is not possible to entirely separate "interface" considerations from "kernal" considerations. (Beeson, 1998)

Other important works in this area are Heeren, Jeuring, and Gerdes (2010) and Prank (2011). In particular, Aplusix focuses on reasoning by equivalence, see Nicaud et al. (2004). One difficulty,

highlighted by Nicaud et al. (2004), with the menu-driven application of rules occurs when a single rule is applicable to different sub-expressions. For example, in the expression $x^4 - x^2 - 9$ the rule $a^2 - b^2 \rightarrow (a - b)(a + b)$ can be applied in two different ways, either matching to $x^4 - x^2$ or to $x^4 - 9$. In this situation it is difficult to select the sub-expression $x^4 - 9$ from $x^4 - x^2 - 9$ without re-ordering some of the terms. More detailed discussion of these issues is given by Beeson (1998) and Beeson (2004).

6 Discussion

This chapter opened by considering the four "patterns of thought" identified by Polya (1962), namely (i) *"the pattern of two loci"*, (ii) *"superposition"*, (iii) *"recursion"* and (iv) the *"Cartesian"* pattern of thought. The Cartesian pattern of problem solving involves *setting up an equation*, then solving it and interpreting any solution. These patterns relate to how to go about solving problems. However, more than half (21/40) of the separate question parts in the examination questions of Section 3 do not relate to a problem at all. Rather they instruct students to undertake a well-rehearsed set of techniques, isolated from any problem. For example, highers paper 1 Q6 asks students to evaluate $\log_6(12) + \frac{1}{3}\log_6(27)$. Technique alone does not correspond to the Cartesian pattern of thought. All but one of the other question parts asked students to set up and solve equations. One question part does relate to recursion, expecting students to recognise the limit of a recurrence relation would result in the fixed point of an equation. Burkhardt (1987), and others, have argued that in a very real sense to students, the subject of mathematics is defined by what we, as a mathematics community, expect students to do in examinations. Currently, school level problem solving and proof is confined to algebraic and calculus methods in which the primary technique is reasoning by equivalence.

Reasoning by equivalence continues to play a central role in mathematics beyond school examinations, and into undergraduate work. As already mentioned, it forms much of the work in symbolic calculus problems. Further, with many proofs reasoning by equivalence is central. For example, in induction to prove results such as $\sum_{k=1}^{n} k^2 = \frac{1}{6}n(n+1)(2n+1)$ the central work in the induction step is reasoning by equivalence. Therefore all technology designed to support proof in teaching should support reasoning by equivalence as a central component.

In this work I have focused on equivalence, rather than on re-write rules. There is an extensive literature on re-write rules, e.g. (Bundy, 1983), and these are commonly taught in school mathematics. The Mathpert system of Beeson (1998) avoids the problem of deciding if a student's step is legitimate by providing a menu of steps from which the student must choose. It is certainly possible to infer some of the steps a student has made, and so some analysis of students' steps is a valuable addition to the reasoning by equivalence described here. However, steps alone will not be sufficient. Indeed, one goal of developing fluency in algebra is to become proficient at combining small individual steps into one. Taken to an extreme, a student might well move from the equation $x^2 + 2x + 1 = 0$ directly to the solution $x = 1$. These equations are equivalent, but whether this single step is acceptable will depend on the teacher and context. The teacher might additionally require other properties in a complete solution. For example, they may require a minimum number of lines of working, or that a particular form (e.g. factored) appears as one line of working.

Perhaps the most significant drawback of focusing on equivalence is the problem that addition, removal and permutation of equivalent intermediate steps will all be accepted by a pure equivalence reasoning engine. Firstly I should note that students do, in fact, sometime backtrack or write unnecessary steps. While this might well be sub-optimal or less aesthetically pleasing this does not make their work "wrong". Further work with live students will be needed to decide the extent to which this is actually a problem, and to design mechanisms for avoiding the problems which arise. Ultimately in addition to equivalence the teacher will seek to establish a range of other properties which may include inferring which steps were taken where this is possible, and the relationship (if any) to a teacher's envisaged model answer.

Current dynamic geometry systems provide powerful tools for experimentation. However, none of the current systems of which I'm aware allow students to move beyond experimentation to write up a more formal claim and proof. Allowing students to write simple proofs based on Polya's *"pattern of two loci"* appears to be a sensible starting point, and combining algebra and geometry another.

There are more profound differences at the level of logic. Most automatic theorem proving software makes use of the resolution rule of inference which is closely related to the contrapositive, Bundy (2013). To use existing automatic theorem proving software a more profound shift in how we teach logic is needed.

7 Conclusion

This chapter takes the hypotheses that students continue to need to learn how to perform reasoning by equivalence accurately, regardless of the technology available to them. More specifically, the hypothesis is that all students will need to encounter some reasoning by equivalence as a basic part of algebra: teaching how to solve linear and quadratic equations is unlikely to disappear from curricula in the future. In the version of STACK described here, students are expected to perform the computations themselves and type in their answer, hence the need for flexibility of what is and is not a step. In contrast, when using MathXpert students chose what to do, but were not expected to do it. Not saying what you are doing is defensible when only reasoning by equivalence. However, as soon as other operations are included (e.g. let, equate coefficients, and calculus) then this position becomes unsustainable. One solution is what Back, Mannila, and Wallin (2010) called *structured derivations*. Another very early approach, illustrated in Brancker, Pell, and Rahn (1668), is the three column proof in which the argumentation and reasoning are clearly laid out with reference to numbered lines.

We do have an opportunity, through carefully designed tools, to communicate to students what is and is not important in writing an acceptable mathematical proof. If logical symbols such as $\Leftrightarrow$ matter, then either the system should include them automatically or students should be constrained to write them. Templates provide another alternative in creating constraints in which people work. Constraints can be liberating: they remove the need to worry about whether the overall *form* of the proof is acceptable, leaving the proof author instead to focus on the details. Designing software to automate a process is another, very demanding, type of constraint. I think the attempt to automate assessment of students' proofs is a valuable way of understanding the nature of elementary mathematics, and the challenges associated with teaching and learning mathematics.

References

Adams, W. W., & Loustaunau, P. (1994). *An introduction to Grobner bases* (Vol. 3). Providence, Rhode Island: American Mathematical Society.

Back, R. J., Mannila, L., & Wallin, S. (2010). 'it takes me longer, but I understand better' – student feedback on structured derivations. *International Journal of Mathematical Education in Science and Technology*, *41*(5), 575–593. doi: 10.1080/00207391003605221

Beeson, M. (1998). Design principles of mathpert: Software to support education in algebra and calculus. In N. Kajler (Ed.), *Computer-human interaction in symbolic computation* (pp. 89–115). Vienna, Austria: Springer-Verlag. doi: 10.1007/978-3-7091-6461-7

Beeson, M. (2004). The mechanization of mathematics. In C. Teuscher (Ed.), *Alan Turing: Life and legacy of a great thinker* (pp. 77–134). Berlin, Germany: Springer. doi: 10.1007/978-3-662-05642-4

Bernardo, G., & Carmen, B. (2009). The ambiguity of the sign $\surd$. In *Proceedings of CERME6, Lyon, France* (Vol. Working Group 4, pp. 509–518).

Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning students use. *Educational Studies in Mathematics*, *75*(1), 89–105. doi: 10.1007/s10649-010-9242-9

Bonnycastle, J. F. (1836). *An introduction to algebra* (16th ed.). London, UK: Longman.

Boole, G. (1847). *The mathematical analysis of logic, being an essay towards a calculus of deductive reasoning*. Cambridge, UK: MacMillian, Barclay, & MacMillan.

Brancker, T., Pell, J., & Rahn, J. H. (1668). *An introduction to algebra*. London, UK: Printed by W.G. for Moses Pitt.

Bundy, A. (1983). *The computer modelling of mathematical reasoning*. London, UK: Academic Press.

Bundy, A. (2013). The interaction of representation and reasoning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *469*(2157). doi: 10.1098/rspa.2013.0194

Burkhardt, H. (1987). What you test is what you get. In I. Wirszup & R. Streit (Eds.), *The dynamics of curriculum change in developments in school mathematics worldwide.* University of Chicago School Mathematics Project.

Durell, C. V. (1930). *New algebra for schools (3 vols)*. London, UK: Bell & Sons.

Euler, L. (1822). *Elements of algebra* (3rd ed.). London, UK: Longman, Hurst, Rees, Orme and Co. (Translated from the French, with the notes of M. Bernoulli and the Additions of M. de La Grange by Hewlett, J.)

Heeren, B., Jeuring, J., & Gerdes, A. (2010). Specifying rewrite strategies for interactive exercises. *Mathematics in computer science*, *3*(3), 349–370. doi: 10.1007/s11786-010-0027-4

Kirshner, D., & Awtry, T. (2004, July). Visual salience of algebraic transformations. *Journal for Research in Mathematics Education*, *35*(4), 224–257. doi: 10.2307/30034809

Leibniz, G. (1966). *Logical papers: a selection*. Oxford, UK: Oxford University Press.

Levenson, E. (2012, June). Teachers' knowledge of the nature of definitions: The case of the zero exponent. *The Journal of Mathematical Behavior*, *31*(2), 209–219. doi: 10.1016/j.jmathb.2011.12.006

Lund, T. (1852). *The elements of algebra designed for the use of students in the university* (14th ed.). London, UK: Longman, Brown, Green and Longmans.

Maxwell, E. A. (1959). *Fallacies in mathematics*. Cambridge, UK: Cambridge University Press.

Newman, M. H. A., & et.al. (1957). *The teaching of algebra in sixth forms: a report prepared for the Mathematical Association*. London, UK: G. Bell and sons, Ltd.

Nicaud, J. F., Bouhineau, D., & Chaachoua, H. (2004). Mixing microworlds and CAS features in building computer systems that help students learn algebra. *International Journal of Computers for Mathematical Learning*, *9*(2), 169–211. doi: 10.1023/B:IJCO.0000040890.20374.37

Northrop, E. P. (1945). *Riddles in mathematics: A book of paradoxes*. London, UK: The English Universities Press.

Polya, G. (1962). *Mathematical discovery: on understanding, learning, and teaching problem solving*. London, UK: Wiley.

Prank, R. (2011). What toolbox is necessary for building exercise environments for algebraic transformations. *The Electronic Journal of Mathematics and Technology*, *5*(3).

Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford, UK: Oxford University Press.

Sangwin, C. J. (2015, July). An audited elementary algebra. *The Mathematical Gazette*, *99*(545), 290–297. doi: http://dx.doi.org/10.1017/mag.2015.37

Sangwin, C. J. (2016). Undergraduates' attempts at reasoning by equivalence in elementary algebra. In *Didactics of mathematics in higher education as a scientific discipline: Conference proceedings* (pp. 335–341). Universität Kassel, Leuphana Universität Lneburg, Universität Paderborn.

Sangwin, C. J., & Köcher, N. (2016). Automation of mathematics examinations. *Computers and Education*, *94*, 215–227. doi: 10.1016/j.compedu.2015.11.014

Sangwin, C. J., & Ramsden, P. (2007). Linear syntax for communicating elementary mathematics. *Journal of Symbolic Computation*, *42*(9), 902–934. doi: 10.1016/j.jsc.2007.07.002

Tirosh, D., & Evan, R. (1997). To define or not to define: The case of $(-8)^{\frac{1}{3}}$. *Educational Studies in Mathematics*, *33*(3), 321–330. doi: 10.1023/A:100291660